Assay Validation Methods

To fully assess or validate the performance of a laboratory test, it is important to understand that what we observe, as clinicians, may not always represent the actual disease status of an aquatic population undergoing an epizootic or sub-clinical infection. Apparent prevalence versus true prevalence, and the factors that affect how well these measures correlate is the topic of this section.

Similarly, if we wish to develop a new technique to detect or monitor disease, the first critical step is to determine the objective of the test in terms of sensitivity and specificity, and make sure we select the most appropriate test to meet the objective.

For the majority of diagnostic applications, the test threshold will be established with balance to ensure that multiple objectives are met without disproportionate number of false test results in either direction (false-positives, or false-negatives). In the following sections, we will discuss the difference between apparent prevalence and true prevalence, the factors that affect sensitivity and specificity of a test, and measuring agreement between two tests.

Apparent versus True Prevalence

There is an important difference between apparent prevalence and true prevalence of disease in an aquatic animal population.

Apparent prevalence is the number of animals testing positive by a diagnostic test divided by the total number of fish in the sample tested.

True prevalence is the actual number of diseased animals divided by the numb er of individuals in the population.

Sampling techniques and Quality Assurance measures in the laboratory always strives to produce data that are accurate and bring the values for apparent and true prevalence as close as possible. The better our tests are in terms of sensitivity and specificity, the better our test results will represent the true prevalence of disease in a population. The more accurately we can detect pathogens or measures of abnormal pathology, the more effective we can be at taking necessary steps to manage disease, and control the spread of disease among aquatic populations.

True prevalence may never really be known for a population, unless all animals in a population are tested with an assay that is 100% accurate, which is extre mely rare in human, or veterinarian medicine health tests . For diagnostic testing, or fish health inspections in fish populations, lethal samples are needed to assess disease prevalence. Testing the entire population would be counterproductive, and probably not worth the effort for the additional level of accuracy obtained, when often we are only interested in knowing IF the disease is present. Even when we are presented with a natural or aquaculture disaster, that produces mortality of all animals, the logistics of sampling and testing every fish is costly and daunting to the fish health laboratory.

Because these tests are not 100% accurate, it is important for the fish health specialist to know how to interpret test results using his/her knowledge of how the various tests perform. Given a set of test results, we need to know what proportion is truly positive, and what proportion are false positives. This can be just as important for negative results as well, what proportion are truly negative, and what proportion are false negatives.

Interpreting Diagnostic Tests

Often the best way to understand these measures is to look at a hypothetical situation in which your test is 100% accurate to discern what proportion are expected to be positive and negative, then compare this data to what proportion actually tests positive and negative. First, let's look at the hypothetical 100% accurate test, and the test results we see.

Table 1. A screening test only provides apparent prevalence.

Apparent Prevalence =	**0.1 or 10%**	**DISEASE Status**		
		D + Positive	**D - Negative**	
TEST Results	**Test +**			**100**
	Test -$^{-}$			**900**
		???	**???**	**1000**

When we are only presented with apparent prevalence, we have to make decisions based on this data. We often want to know what proportion of the individuals is expected to true-positives and what proportion will be false positive? Similarly, what proportion of the test negatives will be true- negatives versus false negatives?

Predictive Value of a Positive Test gives the proportion of the test-positives which are true-positives to the total number testing positive. In this table, 9 of the 100 reported positive tests represent truly-positive individuals, giving a low P.V. of 9%. We see the rate of false positives is high (as seen in the 91 disease-negative samples that tested positive).

Table 2. Predictive Values of a Positive or Negative Test.

Apparent Prevalence =	**10%**	**DISEASE Status**			
True Prevalence =	**1%**	**D + Positive**	**D - Negative**		
TEST Results	**Test +**	**9**	**91**	**100**	**P.V.(+) = 9 %**
	Test -	**1**	**899**	**900**	**P.V.(-) = 99.8 %**
		10	**990**	**1000**	

Predictive Value of a Negative Test gives the proportion of the test-negatives which are true-negatives to the total number testing negative. In the example, 899 of the 900 negative test results are correct, giving a P.V.– of 99.8%. Here, the rate of false negatives is low (less than 1%) for this test.

Predictive values are influenced by the sensitivity, specificity, and disease prevalence. PVs should not be confused with sensitivity and specificity. For example, in the above sample set, 9 of the 10 true-positive fish are detected, giving a test sensitivity of 90%. Alternately, 899 of the 990 true-negative fish are reported as negative, giving a specificity of 90%.

Both of these values are very respectable for screening tests. We'll review sensitivity and specificity, and then address how prevalence influences Predictive Values.

Sensitivity and Specificity

SENSITIVITY is the proportion of true-positives which actually test positive, and how well a test is able to detect positive individuals in a population. A sensitive test will rarely "miss" positive individuals, and should be used when the chance of missing disease poses a large penalty (i.e., introduces a serious or exotic disease).

SPECIFICITY is the proportion of true-negatives which actually test negative, and reflects how well an assay performs in a group of disease negative individuals. A specific test will not produce false positives, or misclassify the identity of a pathogen. A highly specific test should be employed when false-positive results would cause significant impacts to the program (i.e., erroneous reporting of a significant disease in humans, or eliminating rare animals from a broodstock program).

SENSITIVITY = $\mathbf{\dfrac{a}{a + c}}$ **SPECIFICITY =** $\mathbf{\dfrac{d}{b + d}}$

Table 2. Sensitivity and Specificity formula.

		DISEASE Status	
		D+	**D -**
TEST Results	**Test +**	**A**	
	Test -	**C**	
		a + c	

	DISEASE Status	
	D+	**D -**
Test +		**b**
Test -		**d**
		b + d

It should be noted that the value for specificity is harder to calculate since defining a disease-free individual is more difficult than detecting a disease-positive individual.

The ideal test is of course both highly sensitive and highly specific. However, the majority of diagnostic tests have continuous outcomes (i.e., quantity of protein, antibody titer, etc.), with the assignment of a positive/negative cut-off value decided by the diagnostician, or authoritative body conducting the testing. This setting of the positive/negative threshold makes these tests appear to produce dichotomous outcomes (i.e., positive or negative).

What occurs when sensitivity or specificity is altered?

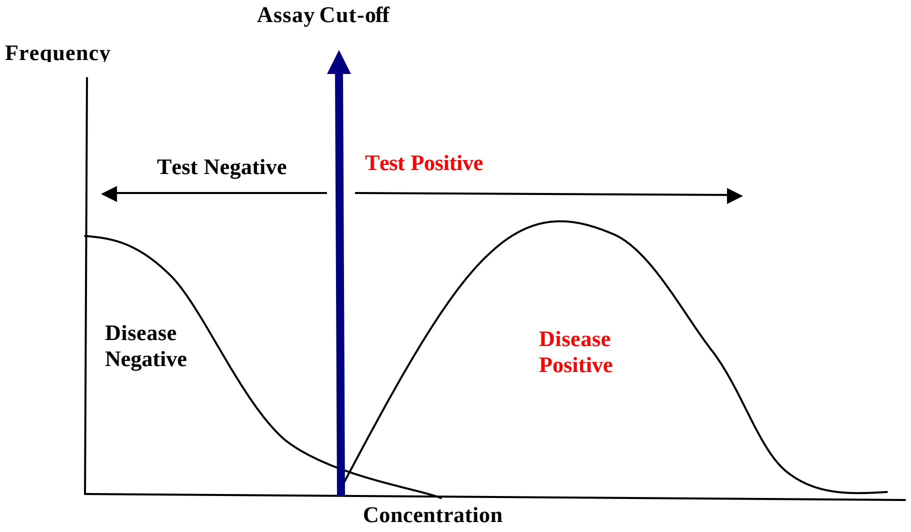
When either test sensitivity or specificity is altered, it will almost always cause a corresponding change in the other. For example, if in a given test, if the cut-off threshold is set lower, to ensure inclusion of all positive fish, it will also include a larger proportion of false-positive test results.

Low Threshold:

All disease-positive fish are included in Positive Test Results:

However, a proportion of disease-negative fish will also be reported in the Positive Test Results;

FALSE POSITIVES



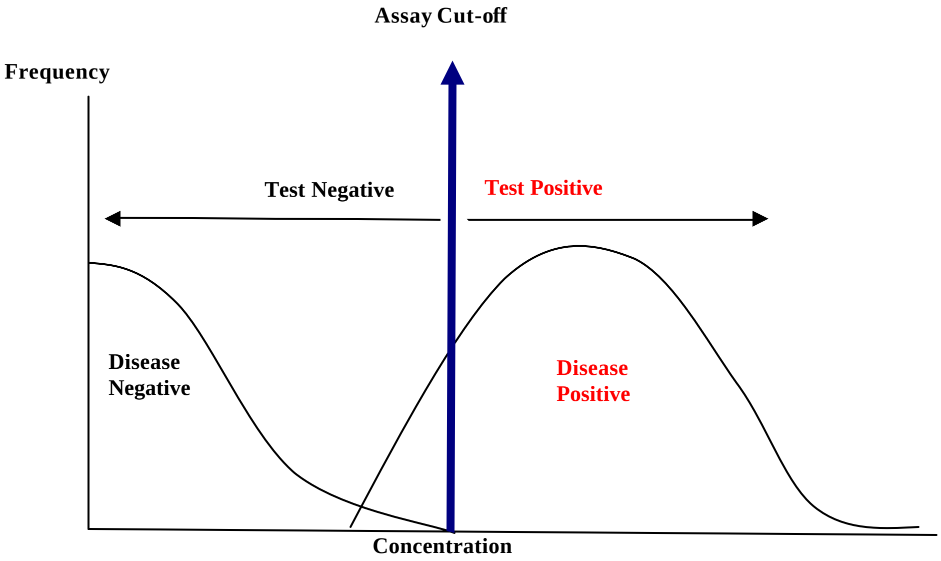
On the other hand, if the threshold is set very high to exclude all negative fish in the positive test results, then the test will also identify a large proportion of true-positive fish, reported as negative test results (false-negatives).

High Threshold:

All disease-negative fish are included in the Negative Test Results:

However, a proportion of disease-positive fish will also be reported in the Negative Test Results;

FALSE NEGATIVES



Determining where to set the cut-off value is very important, and often is driven by the level of error one is willing to accept in either direction. For example, let's look at the case of a captive breeding program, where every fish is extremely valuable, and as many eggs as possible are needed for the restoration program (i.e., restoration is a priority over disease status of a single pathogen).

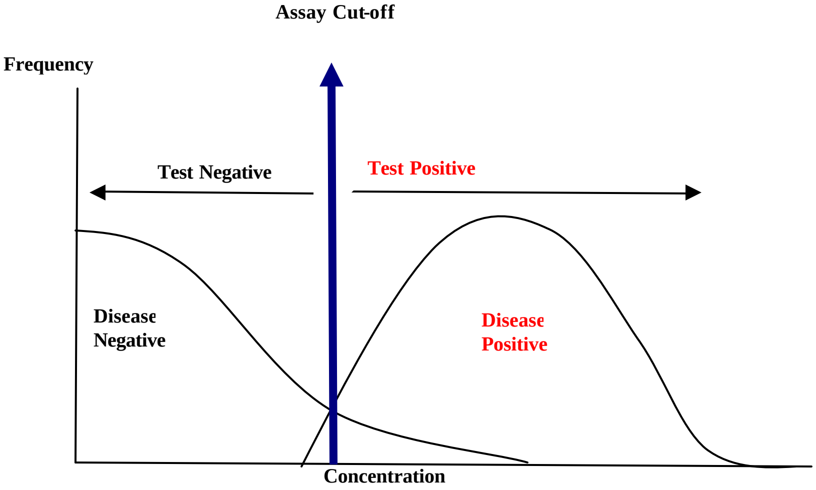
In general, as: Sensitivity increases (lower threshold), false-positive test results will increase,
Specificity increases (higher threshold), false-negative test results increase.

Example: Hatchery practice for salmonid broodstock is to discard all eggs from females testing positive for Bacterial Kidney Disease due to vertical transmission of this disease. In setting up a captive breeding program for recovery of a threatened and endangered (T&E) broodstock, the decision might be made to set the assay threshold higher for a group of fish already at critically low numbers. By setting the positive/negative threshold higher, you would be ensuring that none of the truly-negative fish are discarded as false-positive fish. The downside of this approach would be the acceptance of some truly-positive fish that will test negative (false negatives). But, as stated previously, this may a risk you are willing to take, versus potentially discarding valuable eggs from negative fish. Genetic diversity (maintenance of as many genotypes in the next generation) usually outweighs disease management in these types of scenarios.

When an assay is validated, through testing the sensitivity and specificity using a known positive population, or spiked positive sample set, the threshold can be set in a balanced manner, or to meet specific objectives. A confidence interval can be constructed (see Evaluating Diagnostic Tests) and the clinician will be aware of the test errors and accepted trade-offs in the testing program.

Balanced Threshold:

Small proportions of false positive and false negative test results cannot be entirely avoided in most disease testing. But the degree of error can be understood through validation testing, and the proportions adjusted to meet the objectives of the testing program.



In review, we have covered the following definitions and looked at examples of how sensitivity and specificity are altered when thresholds are established for a diagnostic test.

Table 4. Review of Testing Terms and Formulas.

		DISEASE Status		
		D +	D -	
TEST Results	T +	a	b	a + b
	T -	c	d	c + d
		a + c	b + d	n

Apparent Prevalence	(a+b) / n
True Prevalence	(a+c) / n
Predictive Value of Positive test	a /(a + b)
Predictive Value of Negative test	d /(c + d)
Sensitivity	a /(a + c)
Specificity	d /(b + d)

Let’s return now to the predictive values for positive and negative tests and see how true prevalence in the population affects these values. In this example, the true prevalence is 50%, meaning one would expect that any individual, randomly selected, has a 50% chance of being positive for the disease.

Table 5. Apparent and True Prevalence – Predictive Values of a Test.

True Prevalence (%):	50	DISEASE Status		Apparent Prevalence (%):	50
		D + Positive	D - Negative		% Correct
TEST Results	Test +	475	25	500	95 %
	Test -	25	475	500	95 %
	Total Disease Status (+/-)	500	500	1,000	

Sensitivity	**95**	**%**
Specificity	**95**	**%**

After applying this test, we are 95% sure that test-positive fish truly have the disease, and 95% of the test-negative fish are free of the disease. Also the apparent prevalence and true prevalence are the same. If we use the same screening assay to test a population with a true prevalence of 10%, the predictive values change.

At 10% prevalence, the predictive value of a positive test is significantly lower at 67%.

Table 6. Apparent and True Prevalence (10%) – Predictive Values of a Test.

True Prevalence (%):	10	DISEASE Status		Apparent Prevalence (%):	14
		D +	D -		% Correct
TEST Results	Test +	95	45	140	67.9%
	Test -	5	855	860	99.4%
		100	900	1,000	

Sensitivity	**95**	**%**
Specificity	**95**	**%**

At 1% true prevalence, the predictive value of a positive is 15%. This means that 85% of the test-positive individuals are incorrectly identified.

Table 7. Apparent and True Prevalence (1%) – Predictive Values of a Test.

True Prevalence (%):	**1**	**DISEASE Status**		**Apparent Prevalence (%):**	**5.9**
		D +	**D -**		**% Correct**
TEST Results	**Test +**	**9**	**49.5**	**59**	**15.3%**
	Test -	**1**	**940**	**941**	**99.9%**
		10	**990**	**1,000**	

Sensitivity **90** **%**

Specificity **95** **%**

At 1% true prevalence, the predictive value of a positive is <2%, meaning 98% of the test-positive individuals are incorrectly identified.

Table 8. Apparent and True Prevalence (0.1%) – Predictive Values of a Test.

True Prevalence (%):	**0.1**	**DISEASE Status**		**Apparent Prevalence (%):**	**5.1**
		D +	**D -**		**% Correct**
TEST Results	**Test +**	**1**	**50**	**51**	**1.9%**
	Test -	**0**	**949**	**949**	**100%**
		1	**999**	**1,000**	

Sensitivity **100** **%**

Specificity **95** **%**

In summary, the predictive value decreases as the true prevalence of a population decreases. So, at very low prevalence levels (0.1 -1%), as in carrier or sub-clinical infections, the predictive value of a positive test becomes very small and is not very informative. At higher prevalence levels (i.e., during a severe epizootic), the predictive values will be much stronger.

Comparable Testing Methods – Measuring Agreement

The true disease status is frequently unknown, or impossible to obtain with reasonable effort and costs. In many cases, we use imperfect tests for which there is no quantitative measurement of sensitivity or specificity. Even when we spike sample sets with known pathogens, we cannot be sure these sample sets mimic what occurs in a natural infection.

When new technology provides new methodologies, the new test is often compared to the standard testing methods already in practice. Most frequently, the test producing the greatest number of positives is chosen, and assumed to be the best representative of the actual number of positive individuals in the population. This seems to make sense, but from our previous discussion, we know that if tests produce a disproportionate number of false positives, or false negatives.

Often, when two tests are compared, and the total number positive is similar, say for TEST A and TEST B, it is assumed these are the same individuals testing positive in each test. Often, the positive tests on TEST A may be different individuals than the positive tests on TEST B. When this is the case, it may be difficult determining which disease-positive individuals are testing positive. Another assessment that can be done when comparing two tests is to examine the extent of agreement between the tests, taking into consideration the fact that some individuals will test positive on both tests due to chance alone. Let’s compare a bacterial culture test (CULTURE) to an immunological assay (ELISA) in a hypothetical comparison of two tests. Our population of 1000 is tested and the two tests produce these results:

Table 1. Comparison of two tests and measure of agreement

	Standard Test (ST) →			
New Test (NT) ↓	ST +	ST –	Total ST	ST Apparent Prevalence
NT +	99	501	600	60% (.6)
NT –	1	399	400	
Total NT	100	900	1000 (n)	
NT Apparent Prevalence	10% (.1)			

In this example, the observed agreement is 99 (positives) and 399 (negatives) = 498/1000, or 49.8%. This seems reasonable; however we should take into account the agreement that would occur by chance alone.

The **probability of both tests being positive** is the product of the two apparent prevalences:

0.10 x 0.60 = .06

The **probability of both tests being negative** is the product of 1 minus the two apparent prevalences:

0.4 x 0.9 = .36

The sum of these probabilities is the level of **agreement by chance alone**:

.06 + .36 = 0.42, or 42%

The **agreement beyond chance** is the observed agreement minus the chance agreement:

.498 - .420 = .078, or 7.8%

The **maximum level of agreement beyond chance** is 1minus the chance agreement:

1 - .42 = 0.58, or 58%

The quotient is called *kappa* – the agreement beyond chance divided by the maximum chance agreement:

.078 / 0.58 = .13

No agreement beyond chance gives a kappa of zero, and perfect agreement if 1.
A moderate level of agreement is considered when kappa is greater than 0.4 – 0.5

Testing a Test – Do we ever really have a "Gold Standard"?

Establishing gold standards for diseased and disease free individuals has been the most difficult dilemma in evaluating diagnostic tests for many diseases. For example, a positive bacterial culture rarely produces false positives (with confirmatory testing), however culture is considered lacking for detecting all the true-positives (e.g., bacteria die after removal from fish, in transport, or are overgrown by competing bacteria or fungus). So, the culture method represents a reasonable "gold standard" when it is positive, but it cannot be used as a gold standard if the test is negative (there may be many true-positive fish that do not culture and are reported as negative).

As diagnosticians, we often rely on the best available method, for example when DNA is detected by a DNA probe or PCR assay, it is generally accepted that this test result cannot be wrong when conducted properly. It may be used as a gold standard against which all other assay comparisons are made. However, if DNA testing does not tell us anything about the infection level (e.g., tests results are either positive or negative), this assay may not meet our objective for an appropriate diagnostic test.

Another option for testing validation is to use several available test methods for a disease of interest, and define the true-positives in a relative manner. The major problem here is labor and costs, and the fact that some tests may be testing for different things (i.e., protein, DNA or viable organisms), and sub-clinical infections are more likely to be missed on one or more of the tests. Identification of fish with sub-clinical infections is most often the intent of screening methods, so detection at this level must be addressed by the test methods.

Often the term "gold standard" is applied inappropriately, for only tests which measure a highly specific component, and have been validated quantitatively would meet the standard. Often, tests are developed, demonstrate usefulness in detecting a target pathogen at some infection level, and become accepted by the scientific community as "valid methods".

Screening versus Diagnostic Tests

It is important to recognize if the animals being sampled truly represent the disease-positive and disease-free state of the population, otherwise a test may be perceived to be more useful than it is. During validation of a test method, it is critical to test animals that represent the level of infection for which the assay will be applied in the field setting. A screening test for sub-clinical infections will not have the same function as a diagnostic test for clinical disease. Evaluating a new technique in clinically diseased fish may lead to the conclusion that the test is equally effective in detecting low loads of the target pathogen. Be cognizant that there are three groups to consider when representing an aquatic population: infected with gross pathology (clinical signs and/or lesions), infected with no gross pathology, and non infected individuals.

Observer Bias

It is also important that anyone performing a test validation be blinded to the true status of the sample materials. Bias does not occur as a conscious decision to be influenced by previous knowledge; it is a subconscious effort and therefore needs to be controlled for. The extent of agreement between tests, measured by kappa, is often lower when "non-blind" methods of evaluation are used (Martin and Bonnett, 1987). Even prior knowledge of the approximate true prevalence can result in subtle adjustments in the interpretation of the test results. Evaluation of a diagnostic test should have a random order between true positives and true negatives. Preferably, the diagnostic laboratory should not be aware when the evaluation of a test is being performed so that personnel will treat the samples as routinely as possible, avoiding increased attention and special treatment that samples would not normally receive during routine testing.

Sample Size

Evaluation of a diagnostic test to determine its sensitivity and specificity requires the estimation of a proportion or likelihood ratios at each test outcome level (Simel, Samsa, and Matcher, 1991). In theory, a representative sample of 100-200 diseased animals and 2000 or more non-diseased animals should give reasonably precise point estimates for sensitivity and specificity, respectively (Martin, 1984).

Accuracy and Precision

Accuracy refers to the ability of the test to give a true indication of the nature and quantity of the substance or object being measured (Martin, 1977). Accuracy can be low without affecting the sensitivity and specificity (see several examples in the Interpreting a Diagnostic Test section. Also, bear in mind that a test may be 100% accurate, but be of little value if it is measuring a meaningless parameter for a disease of interest.

Evaluation of the accuracy of a test usually is performed by the molecular biologist and is often referred to as the "sensitivity" of the test. Since clinical decisions are often based upon the dichotomous values of a test (negative and positive category), the accuracy is of concern only as one of several influences on sensitivity and specificity. Within limits, accuracy is less important than precision in terms of screening tests (Martin, Meek, and Willeberg, 1987).

Precision refers to the ability of the test to give consistent results in repeated determinations in the same sample or animal (Martin, 1977). To evaluate precision, duplicate testing of the same fish tissues should be performed by the same laboratory, and between different laboratory staff. There can be poor repeatability of test results between laboratory staff when standardized procedures are not clearly defined and followed. Repeatability between laboratories can also be tested when the persons performing the testing are blinded to the sample status.

References

Martin, S.W. (1977). The evaluation of tests. Can Jour Comp Med 41:19-25.

Martin, S.W. (1984). Estimating disease prevalence and the interpretation of screening test results. Prev Vet Met 2: 463-472

Martin, S.W., and B. Bonnett. (1987). Clinical epidemiology. Can Vet J 28:318-325

Martin, S.W., A.H. Meek, and P. Willeberg (1987). Veterinary Epidemiology – Principles and Methods. Iowa State University Press, Ames, Iowa. 343 p.